Logic Extraction for Explainable AI

Susmit Jha

Computer Science Laboratory

SRI

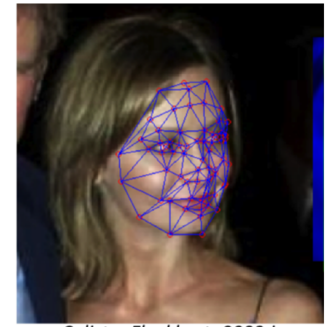
July, 2019

AI reaches human-level accuracy on benchmark datasets

Going deeper with convolutions. (Inception) C Szegedy et al, 2014

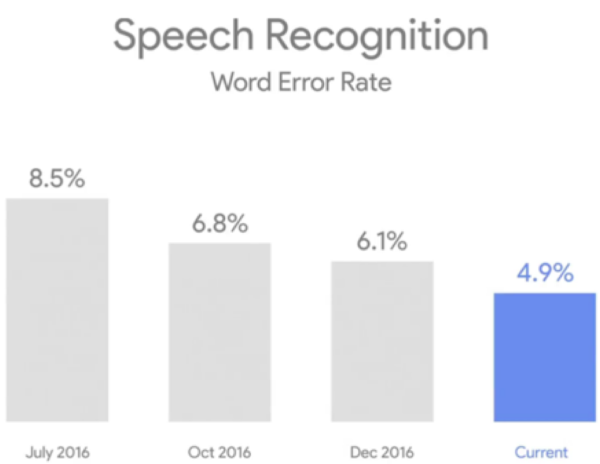
Team	Year	Place	Error (top-5)	Uses external data
SuperVision	2012	1st	16.4%	no
SuperVision	2012	1st	15.3%	Imagenet 22k
Clarifai	2013	1st	11.7%	no
Clarifai	2013	1st	11.2%	Imagenet 22k
MSRA	2014	3rd	7.35%	no
VGG	2014	2nd	7.32%	no
GoogLeNet	2014	1st	6.67%	no

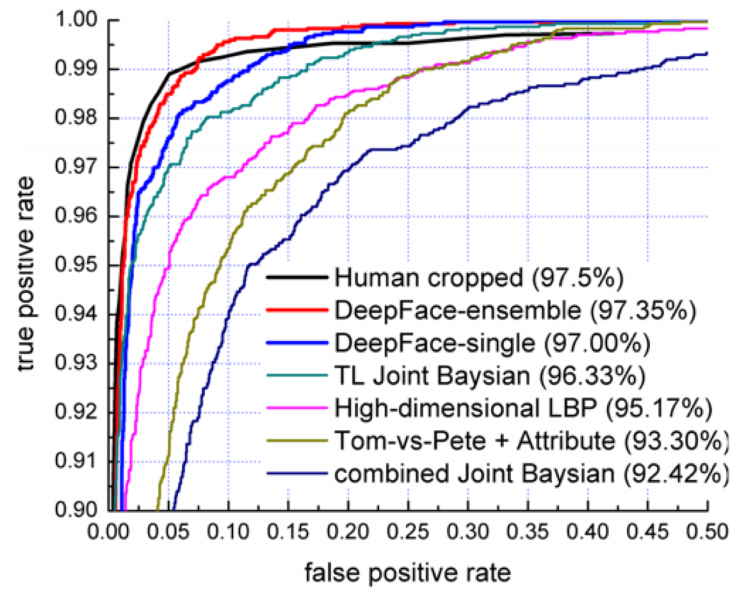
(a) Siberian husky (b) Eskimo dog



Calista_Flockhart_0002.jpg
Detection & Localization

Switchboard benchmark



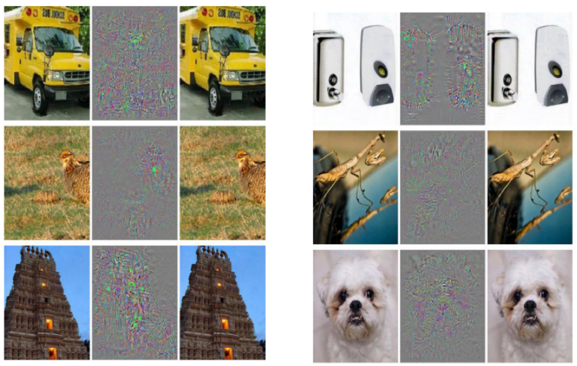


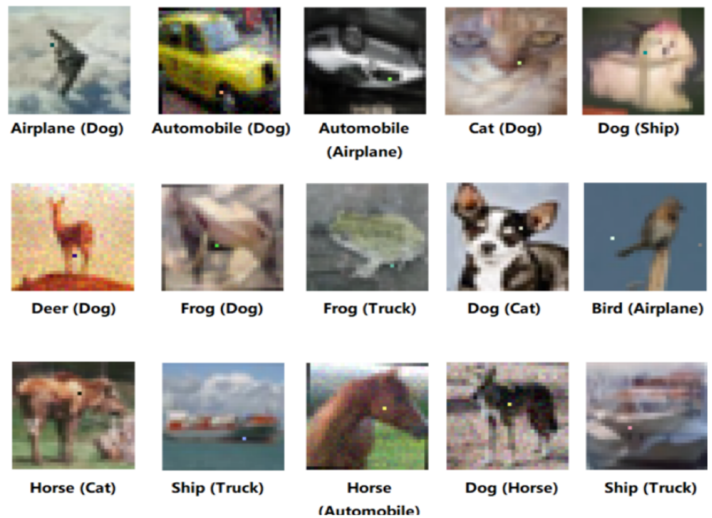
Face Detection. Taigman et al, 2014

Beyond aggregate numbers

Machine learning very susceptible to adversarial attacks.

Szegedy et al, 2013, 2014



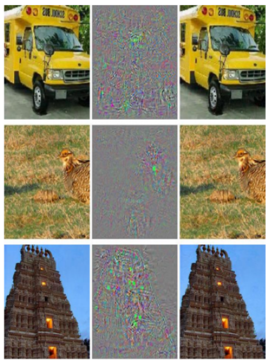


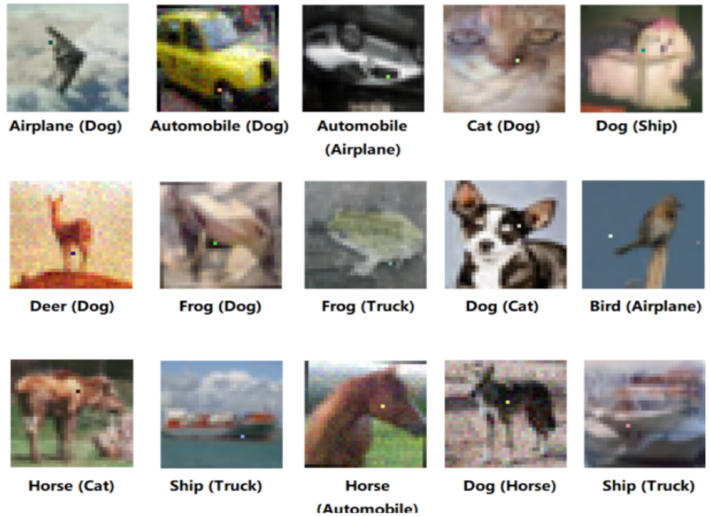
Only allowed to modify the value of 1 pixel. 70.97% of the natural images can be perturbed to at least one target class by modifying just one pixel with 97.47% confidence on average.

Beyond aggregate numbers

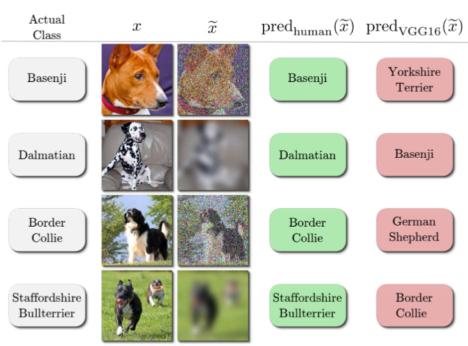
Machine learning very susceptible to adversarial attacks.

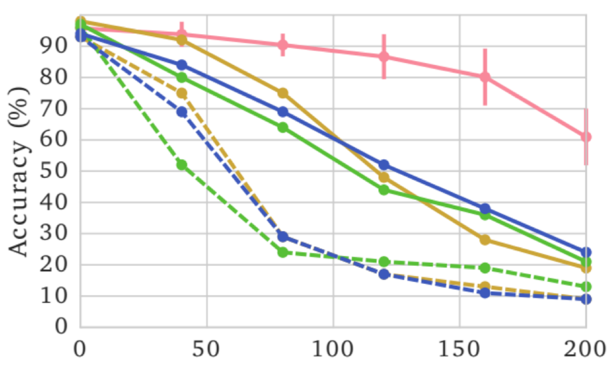
Szegedy et al, 2013, 2014





Only allowed to modify the value of 1 pixel. 70.97% of the natural images can be perturbed to at least one target class by modifying just one pixel with 97.47% confidence on average.



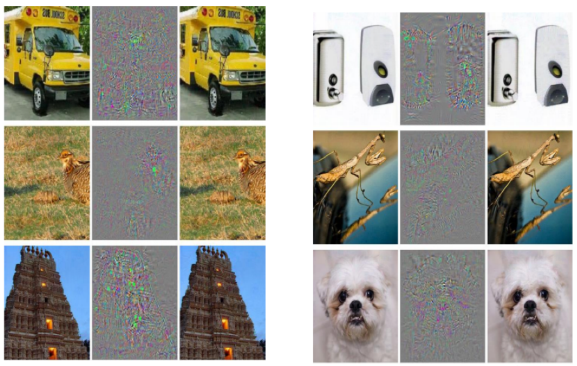


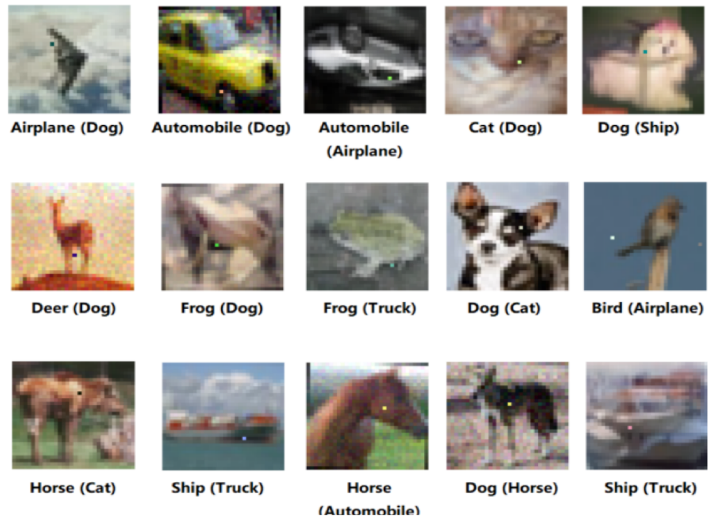
Low robustness to benign noise Dodge et al. 2017

Beyond aggregate numbers

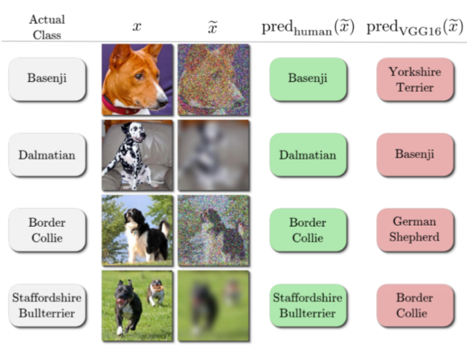
Machine learning very susceptible to adversarial attacks.

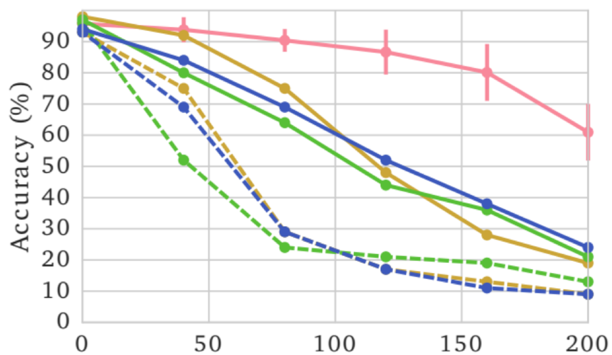
Szegedy et al, 2013, 2014





Only allowed to modify the value of 1 pixel. 70.97% of the natural images can be perturbed to at least one target class by modifying just one pixel with 97.47% confidence on average.





Basketball (73%)

Ping-pong ball (92%)

Statistically good doesn't mean logically/conceptually good.

Low robustness to benign noise Dodge et al. 2017

Understanding deep learning requires rethinking generalization.
C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals

TRINITY: Trust, Resilience and Interpretability

Trust

- Global Assume/Guarantee Contracts on DNNs
- Closed-loop verification of NN controllers
- Extracting and Integrating Temporal Logic into Learned Control

Resilience

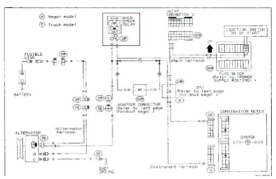
- Adversarial Robustness

Interpretability

- Explaining Decisions as Sparse Boolean Formula Learning
- Inverse Reinforcement Learning of Temporal Specifications

TRINITY: Trust, Resilience and Interpretability

Demonstrations

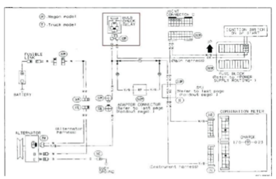


Specifications

TRINITY: Trust, Resilience and Interpretability

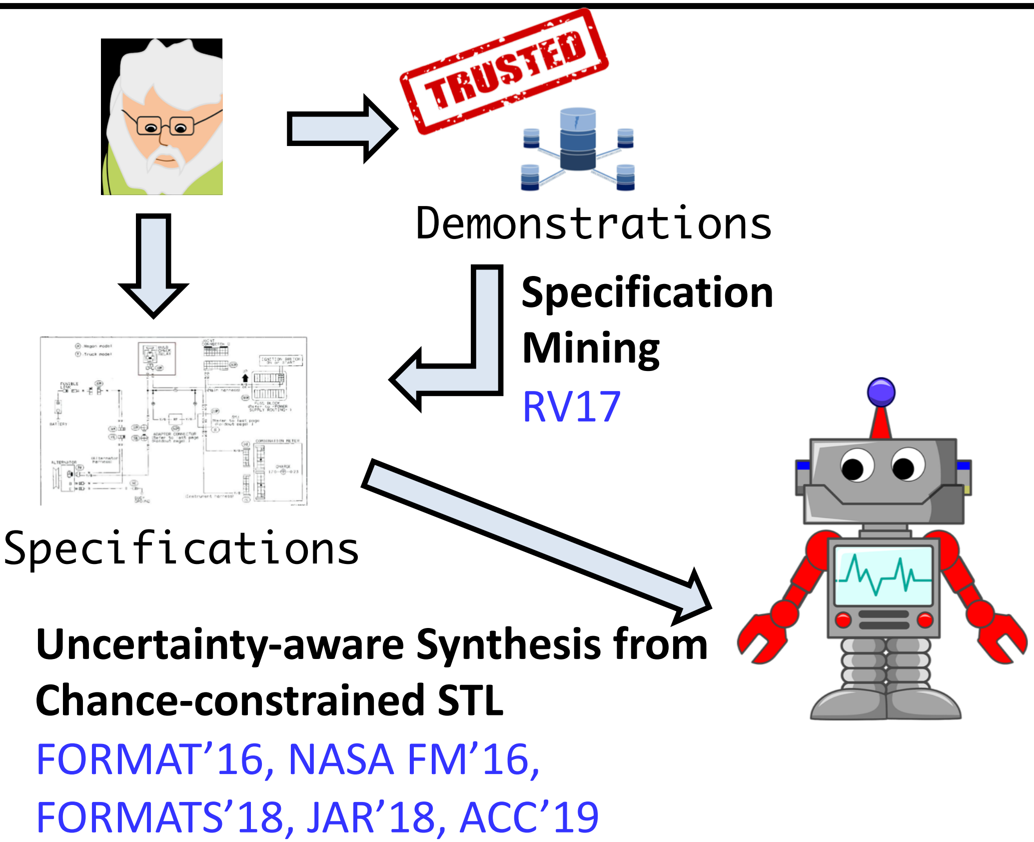
Demonstrations

Specification Mining
RV17

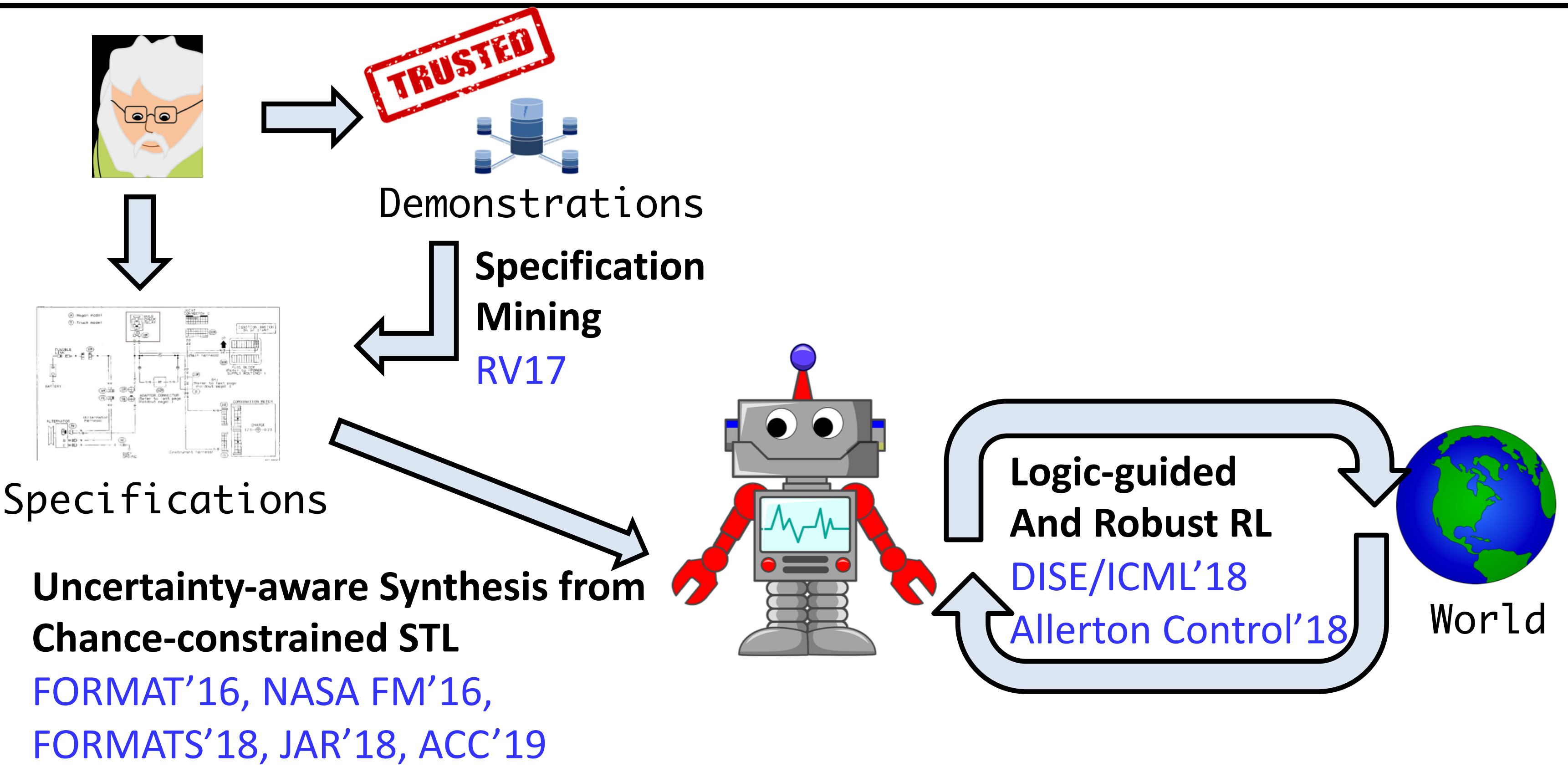


Specifications

TRINITY: Trust, Resilience and Interpretability



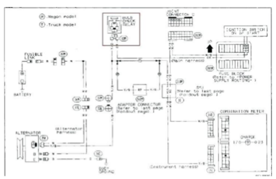
TRINITY: Trust, Resilience and Interpretability



TRINITY: Trust, Resilience and Interpretability

Demonstrations

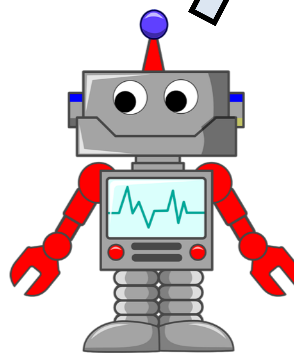
Specification Mining
RV17



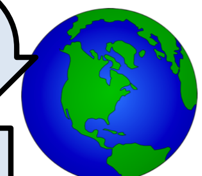
Specifications

Uncertainty-aware Synthesis from Chance-constrained STL
FORMAT’16, NASA FM’16, FORMATS’18, JAR’18, ACC’19

Verification: ML model + closed loop
NASA FM’18, ADHS’18, HSCC’19, VNN/AAAI’19

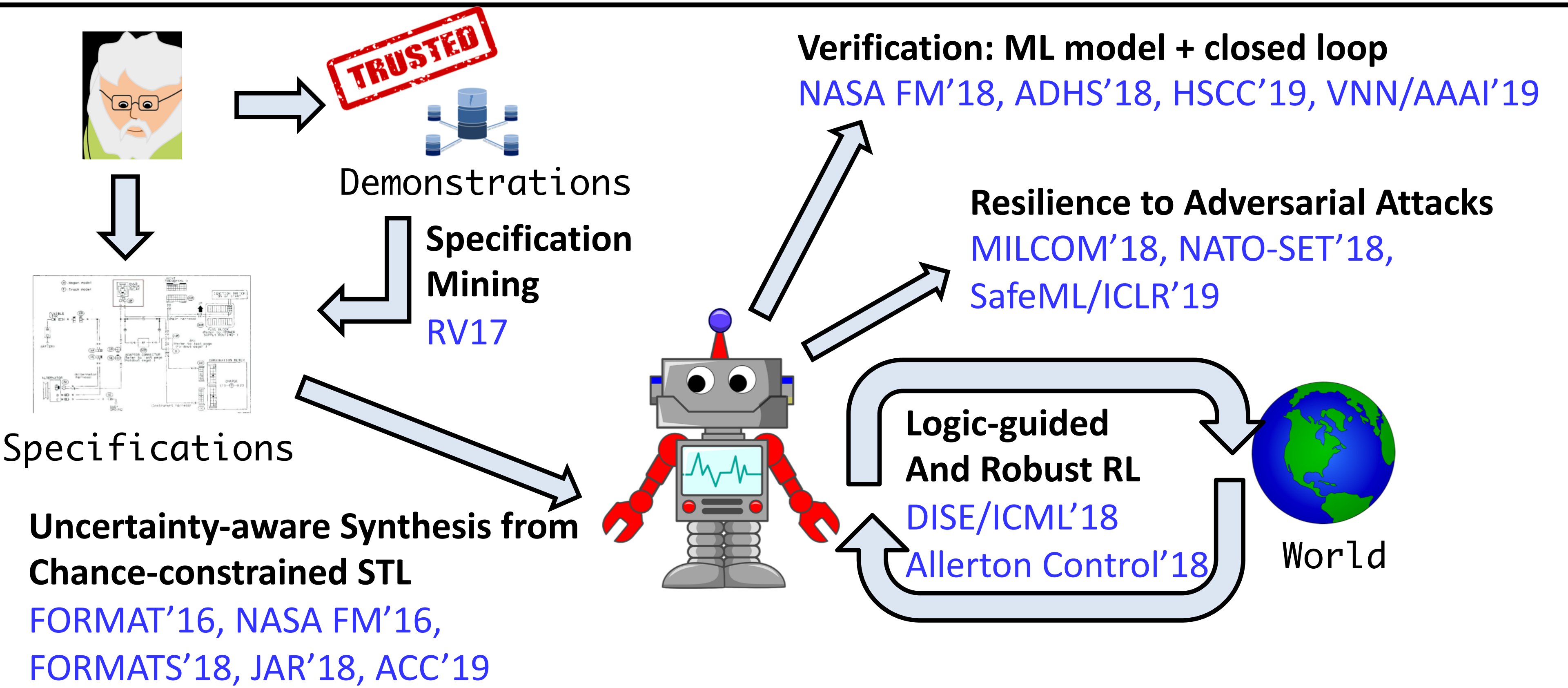


Logic-guided And Robust RL
DISE/ICML’18
Allerton Control’18

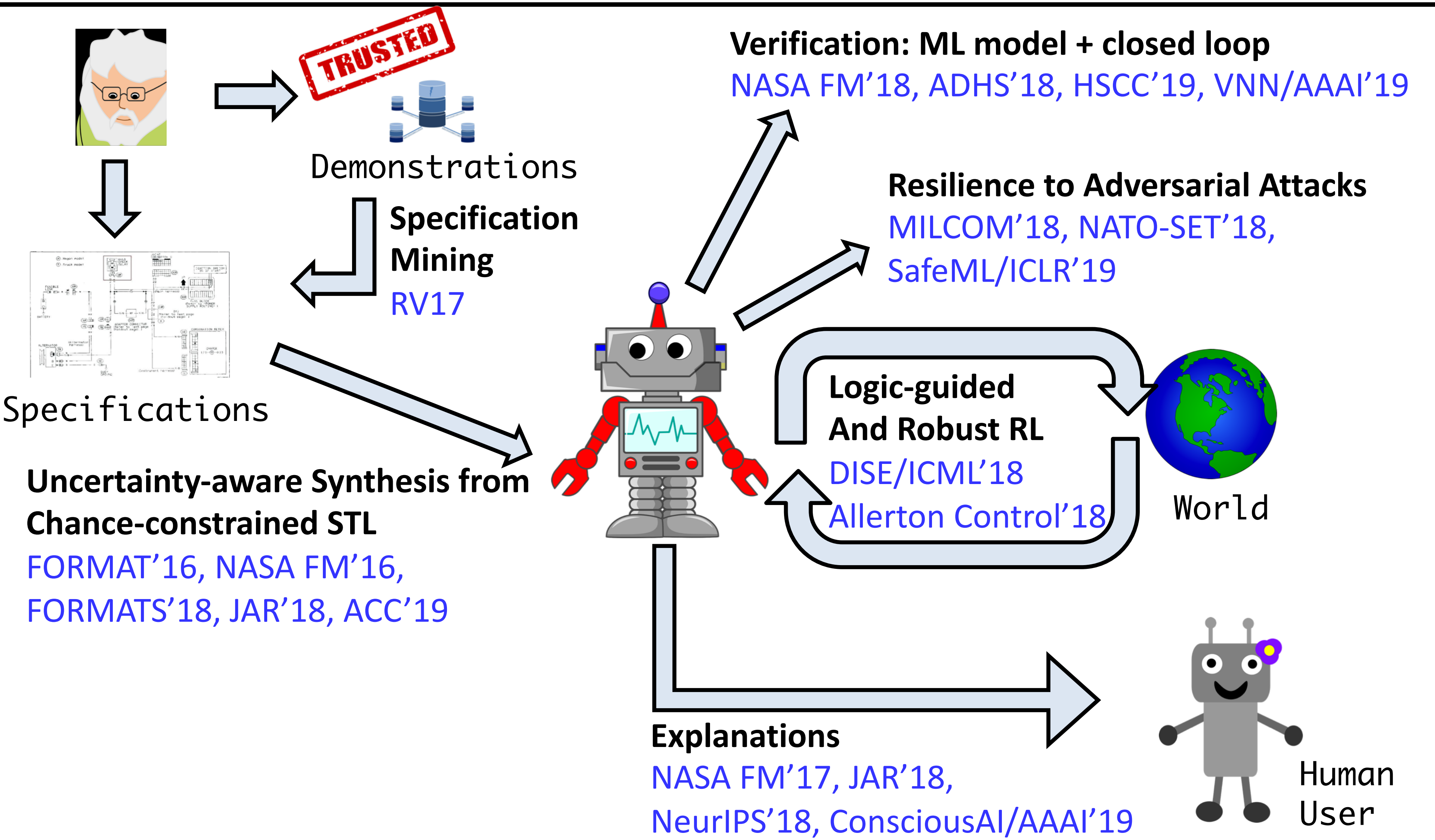


World

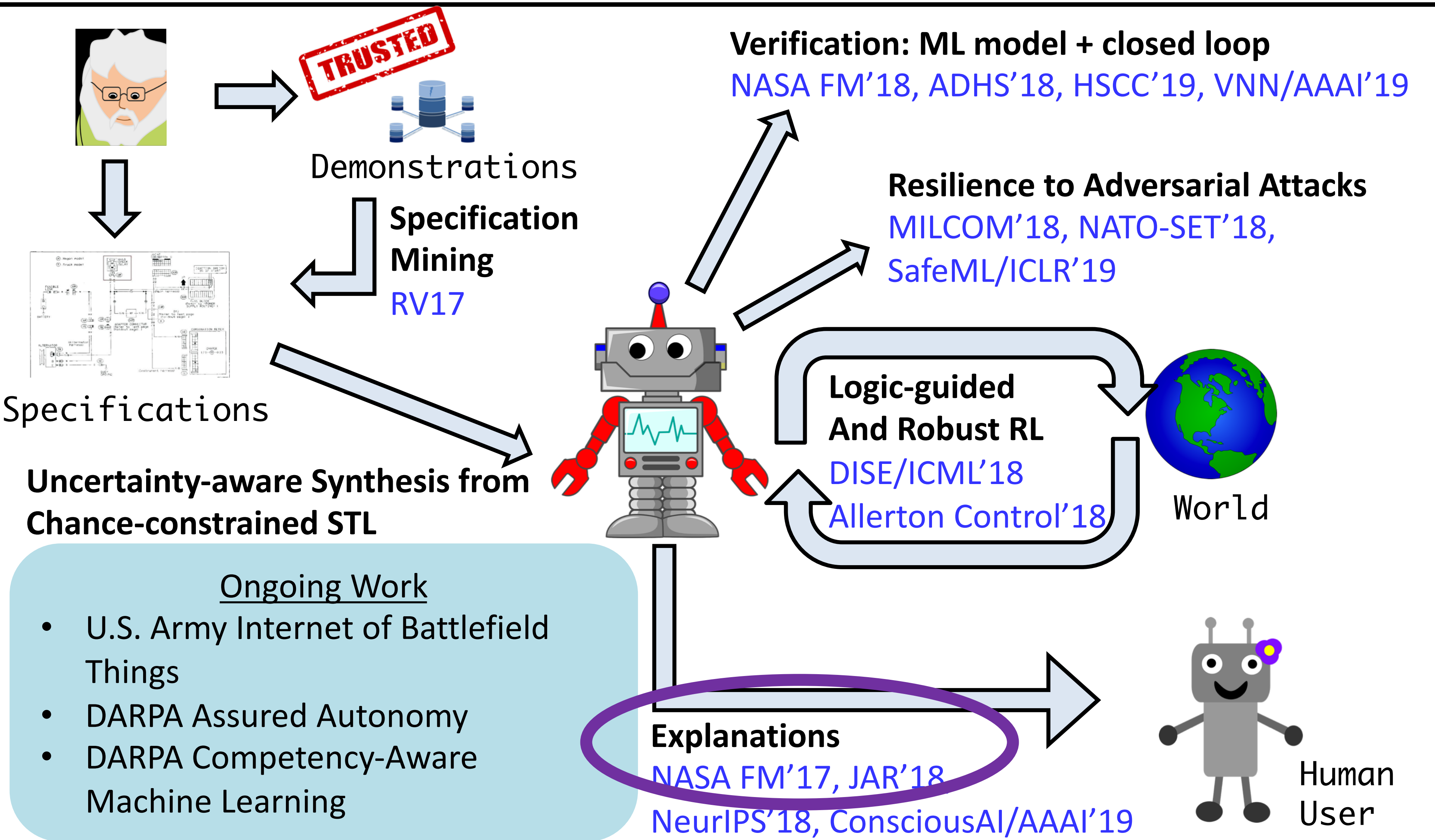
TRINITY: Trust, Resilience and Interpretability



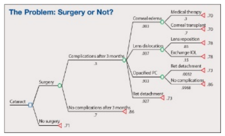
TRINITY: Trust, Resilience and Interpretability

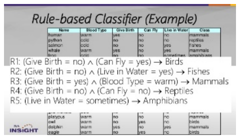


TRINITY: Trust, Resilience and Interpretability

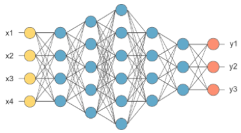


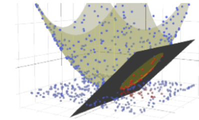
Need for explanation



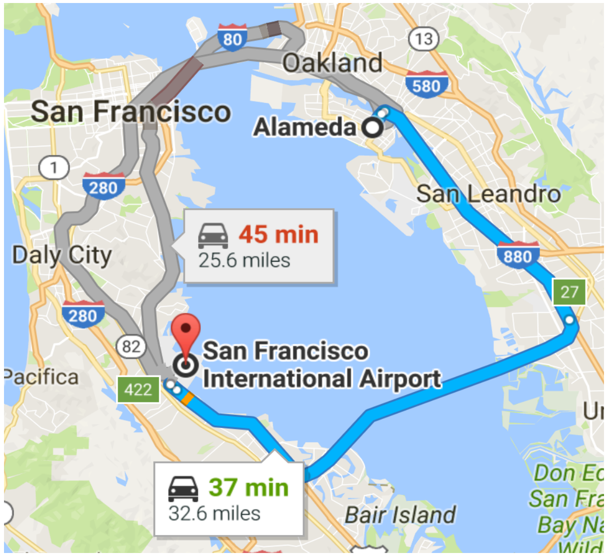


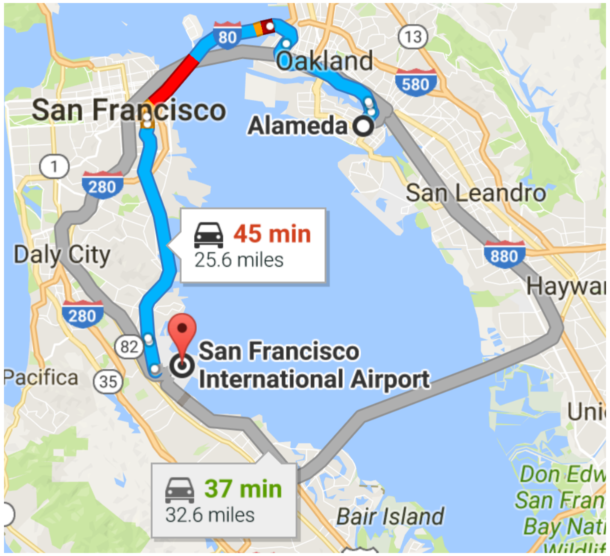
Interpretable but less scalable:
Decision Trees, Linear Regression



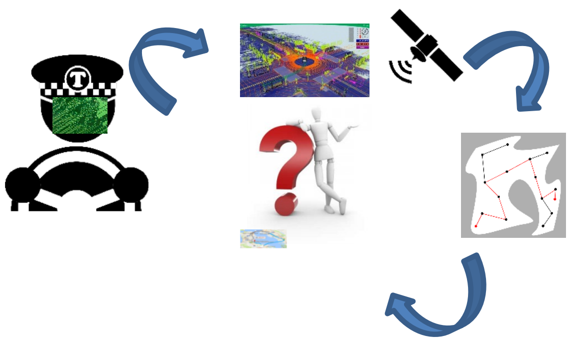


Scalable but less interpretable :
Neural Networks, Support Vector Machines





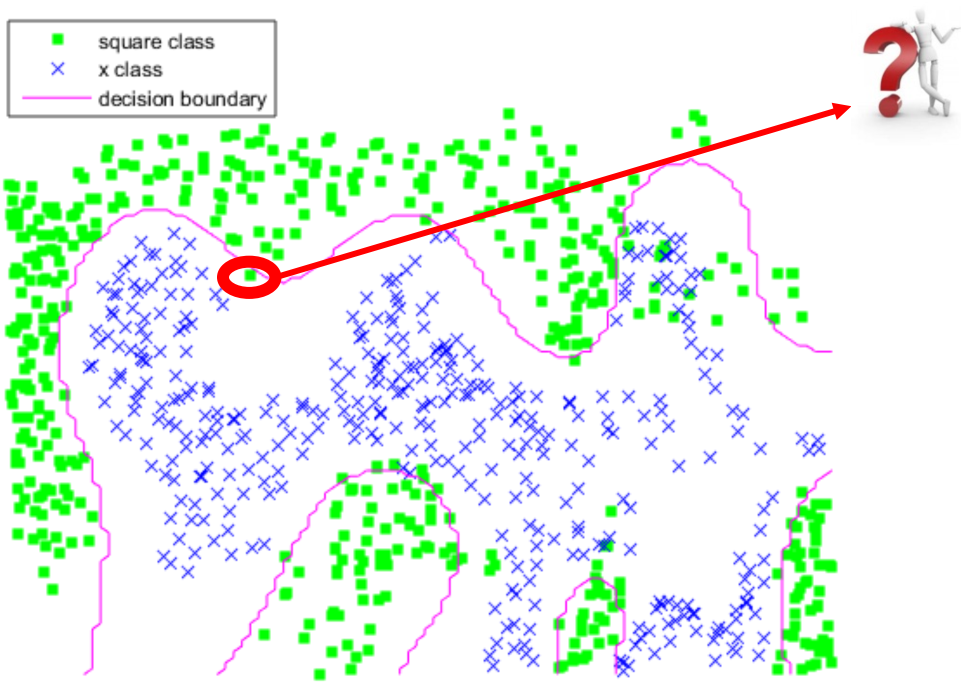
- This route is faster.
- There is traffic on Bay Bridge.
- There is an accident just after Bay Bridge backing up traffic.



Why did we take the San Mateo bridge instead of the Bay Bridge ?

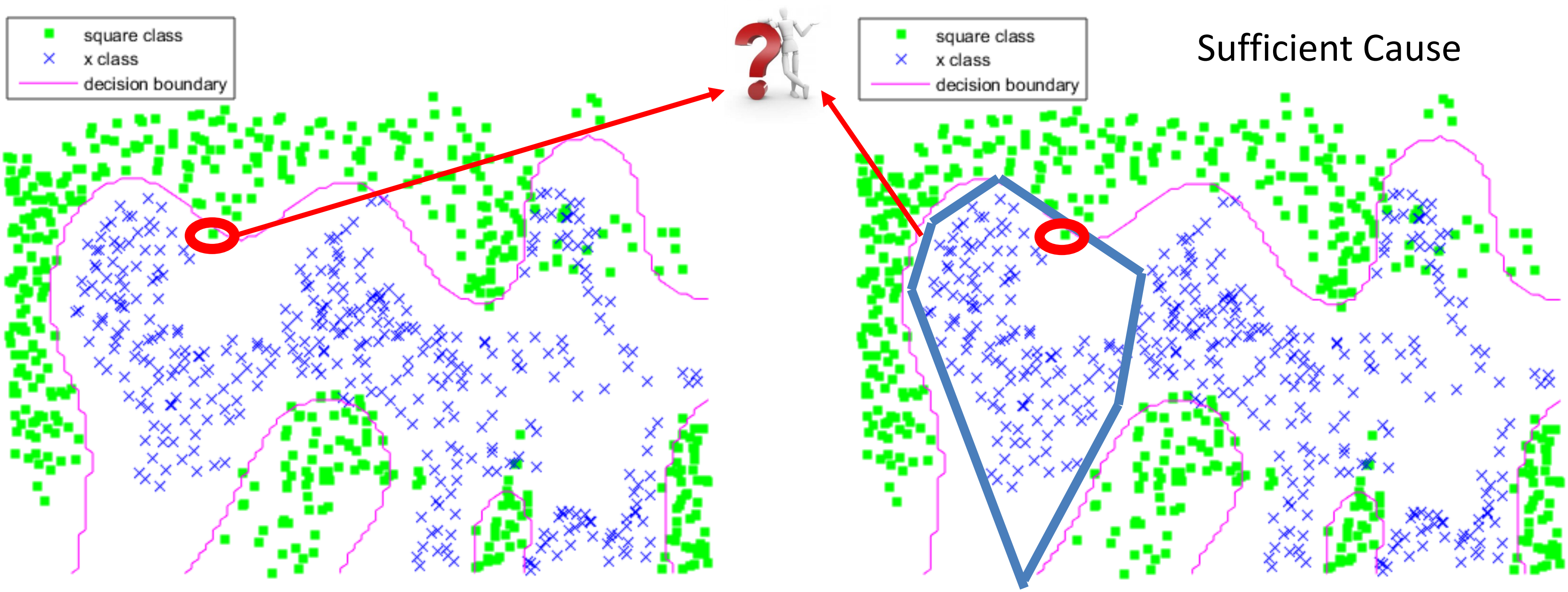
Local Explanations of Complex Models

Not reverse engineering an ML model but finding explanation locally for one decision.



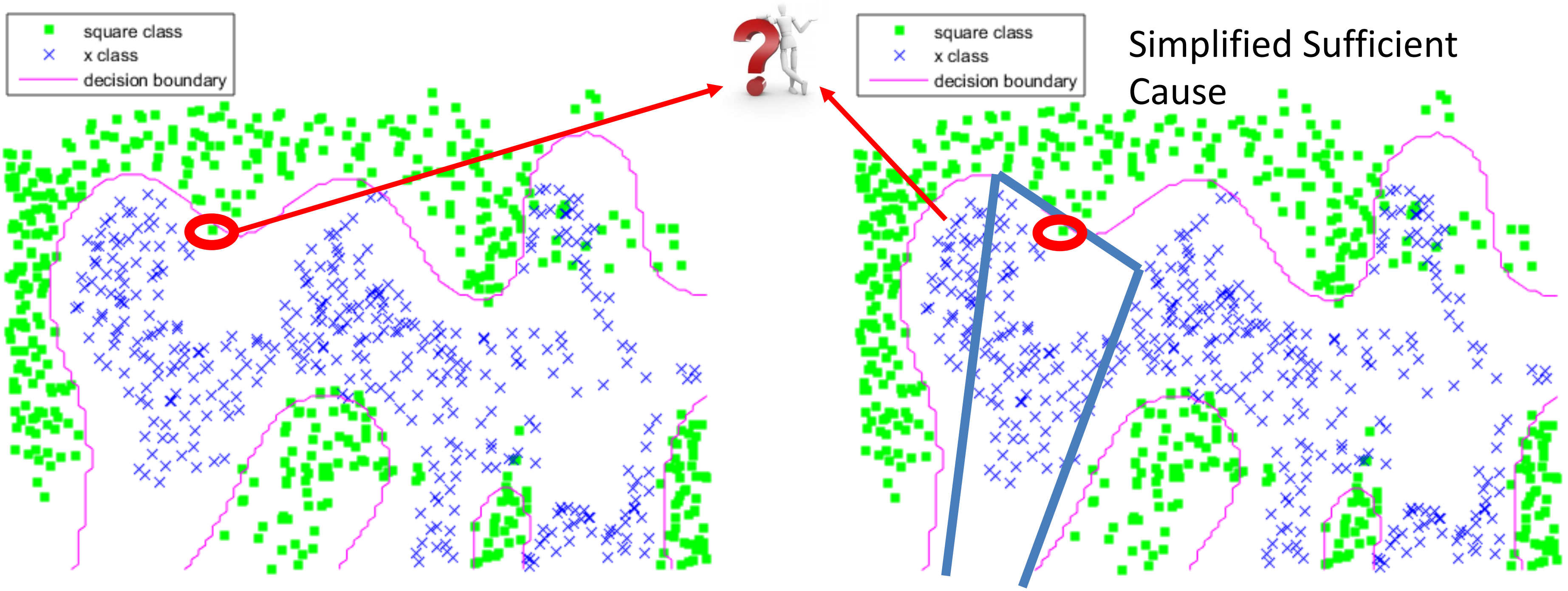
Local Explanations of Complex Models

Not reverse engineering an ML model but finding explanation locally for one decision.



Local Explanations of Complex Models

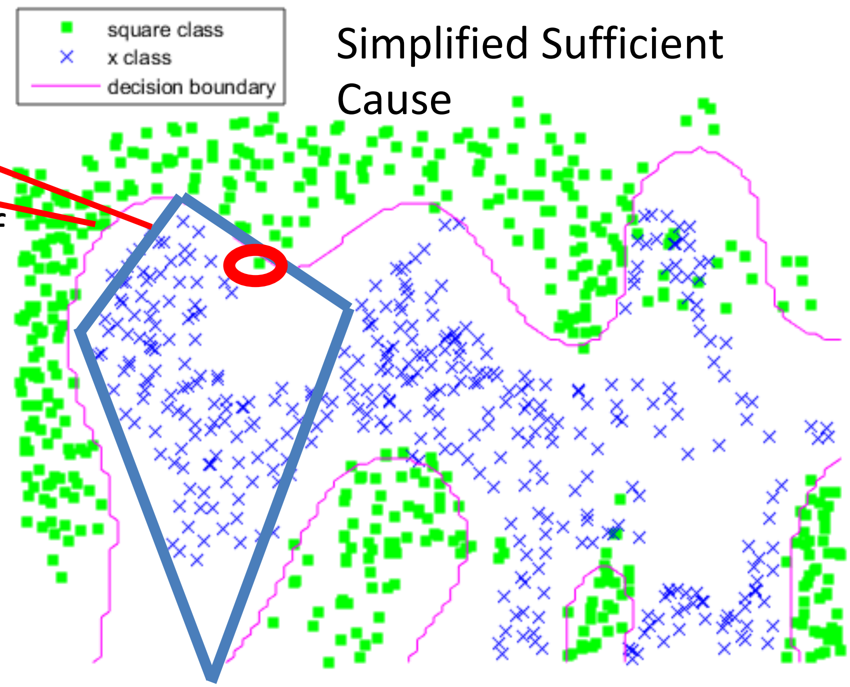
Not reverse engineering an ML model but finding explanation locally for one decision.



Local Explanations in AI

Not reverse engineering an ML model but finding explanation locally for one decision.

$$\xi(x) = \operatorname*{argmin}_{g \in G} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$



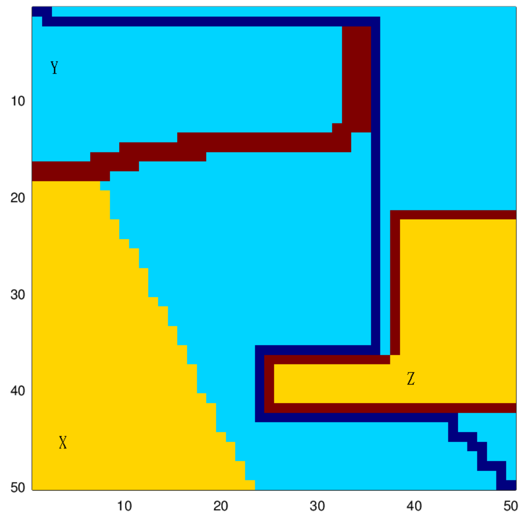
$\mathcal{L}(f, g, \pi_x)$ Measure of how well g approximates f

$\Omega(g)$ Measure of complexity of g

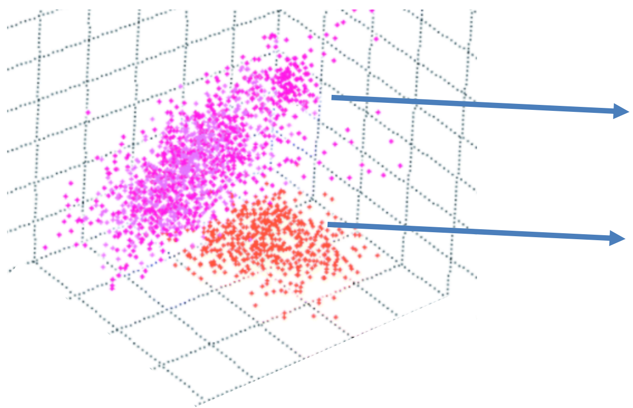
Formulation in AI:

- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?: Explaining the Predictions of Any Classifier." *International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.
- Hayes, Bradley, and Julie A. Shah. "Improving Robot Controller Transparency Through Autonomous Policy Explanation." *International Conference on Human-Robot Interaction*. ACM, 2017.

Model Agnostic Explanation through Boolean Learning



Why does the path not go through Green?



Let each point in k-dimensions (for some k) correspond to a map.

Maps in which optimum path goes via green

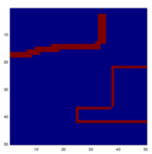
Maps in which optimum path does not go via green

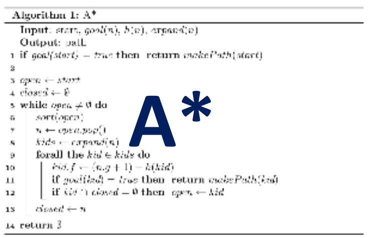
Find a Boolean formula ϕ such that

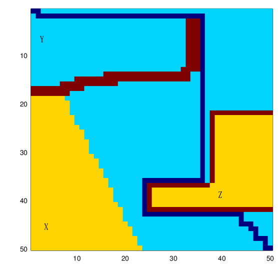
$$\phi \Leftrightarrow Path\ contain\ z$$

$$\phi \Rightarrow Path\ contain\ z$$

Explanations as Learning Boolean Formula







$\phi_{explain}$:
Using explanation vocabulary
Ex: Obstacle presence

ϕ_{query} :
Some property of the output
Ex: Some cells not selected

$$\boldsymbol{\phi_{explain} \Rightarrow \phi_{query}}$$

$$\boldsymbol{\phi_{explain} \Leftrightarrow \phi_{query}}$$

How difficult is it? Boolean formula learning

$$\phi_{explain} \Rightarrow \phi_{query}$$
$$\phi_{explain} \Leftrightarrow \phi_{query}$$

50x50 grid has $2^{2^{50X50}}$ possible explanations even if vocabulary only considers presence/absence of obstacles.

Scalability: Usually the feature space or vocabulary is large. For a map, its order of features in the map. For an image, it is order of the image's resolution.

Guarantee: Is the sampled space of maps enough to generate the explanation with some quantifiable probabilistic guarantee?

How difficult is it? Boolean formula learning

$$\phi_{explain} \Rightarrow \phi_{query}$$
$$\phi_{explain} \Leftrightarrow \phi_{query}$$

50x50 grid has $2^{2^{50X50}}$ possible explanations even if vocabulary only considers presence/absence of obstacles.

Scalability: Usually the feature space or vocabulary is large. For a map, its order of features in the map. For an image, it is order of the image's resolution.

Guarantee: Is the sampled space of maps enough to generate the explanation with some quantifiable probabilistic guarantee?

On PAC learning algorithms for rich Boolean function classes

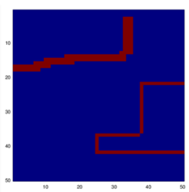
Rocco A. Servedio*

Department of Computer Science
Columbia University
New York, NY U.S.A.
rocco@cs.columbia.edu

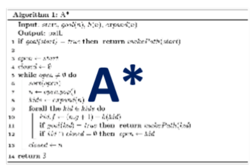
Theoretical Result:

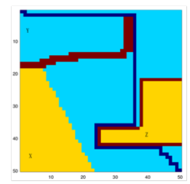
Learning Boolean formula even approximately is hard. 3-DNF is not learnable in Probably Approximately Correct framework unless RP = NP.

Two Key Ideas



$\phi_{explain}$:
Using explanation vocabulary
Ex: Obstacle presence





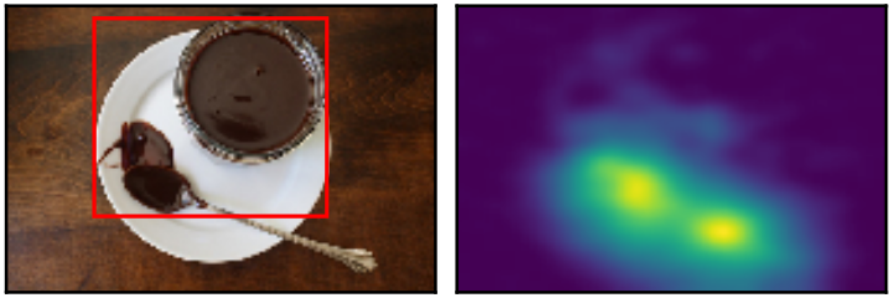
ϕ_{query} :
Some property of the output
Ex: Some cells not selected

1. Vocabulary is large.
2. How many samples (and what distribution) to consider for learning explanation ?
3. Learning Boolean formula with PAC guarantees is hard.

Active learning Boolean formula $\boldsymbol{\phi_{explain}}$ and not learning from fixed sample.

Explanations are often short and involve only few variables !

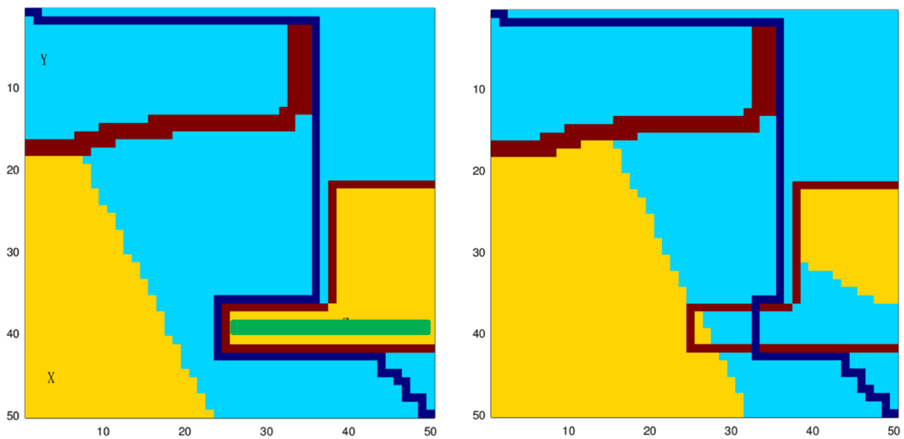
Two Key Ideas



Active learning Boolean formula $\phi_{explain}$ and not learning from fixed sample.

Explanations are often short and involve only few variables !

Two Key Ideas



Involves only two variables.
If we knew which two, we had only $2^{2^2} = 16$ possible explanations.

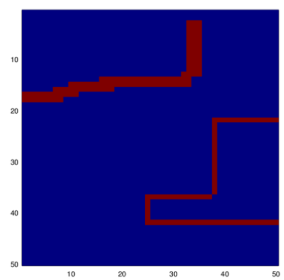
How do we find these relevant variables?

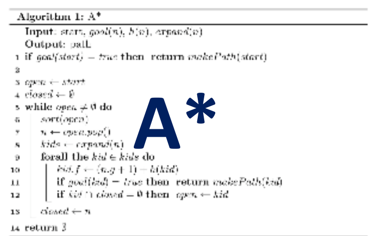
Active learning Boolean formula $\boldsymbol{\phi_{explain}}$ and not learning from fixed sample.

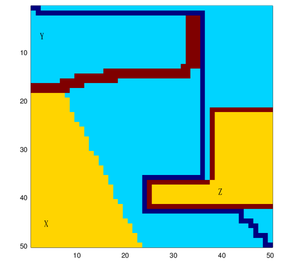
Explanations are often short and involve only few variables !

Actively Learning Boolean Formula

Oracle







ϕ_{query} :
Some property of the output
Ex: Some cells not selected

Assignments to V
m1 = (0,0,0,1,1,0,1)
m2 = (0,0,1,1,0,1,0)

$\phi_{explain}$ (V) :
Using explanation vocabulary
Ex: Obstacle presence

ϕ
Evaluates assignments and returns T,F

Actively Learning Relevant Variables

$$\boxed{Find\ U\ such\ that\ \phi_{explain}(\mathrm{V}) \equiv \phi_{explain}(U)\ where\ |U| \ll |V|}$$

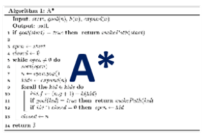
$$\phi_{explain}\ is\ sparse$$

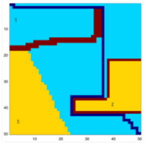
Actively Learning Relevant Variables

$$\boxed{Find\ U\ such\ that\ \phi_{explain}(V) \equiv \phi_{explain}(U)\ where\ |U| \ll |V|}$$

Assignments to V
m1 = (0,0,0,1,1,0,1)

Oracle





ϕ_{query} :
Some property of the output
Ex: Some cells not selected

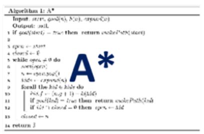
m1 : True

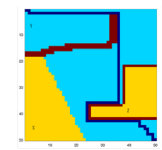
Actively Learning Relevant Variables

$$\boxed{Find\ U\ such\ that\ \phi_{explain}(\mathrm{V}) \equiv \phi_{explain}(U)\ where\ |U| \ll |V|}$$

Assignments to V
m1 = (0,0,0,1,1,0,1)
m2 = (0,0,1,1,0,1,0)

Oracle





ϕ_{query} :
Some property of the output
Ex: Some cells not selected

m1: True, m2: False

Random Sample Till
Oracle differs

Actively Learning Relevant Variables

$$\boxed{Find\ U\ such\ that\ \phi_{explain}(V) \equiv \phi_{explain}(U)\ where\ |U| \ll |V|}$$

Assignments to V

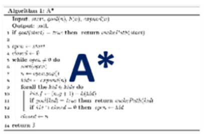
m1 = (0,0,0,1,1,0,1)

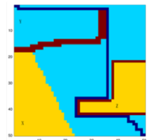
m2 = (0,0,1,1,0,1,0)

m3 = (0,0,0,1,1,1,0)

Binary Search Over Hamming Distance

Oracle





ϕ_{query} :

Some property of the output

Ex: Some cells not selected

m1: True, m2: False

Actively Learning Relevant Variables

$$\boxed{Find\ U\ such\ that\ \phi_{explain}(V) \equiv \phi_{explain}(U)\ where\ |U| \ll |V|}$$

Assignments to V

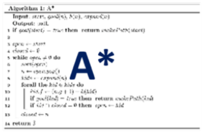
m1 = (0,0,0,1,1,0,1)

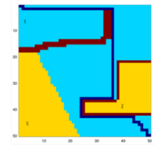
m2 = (0,0,1,1,0,1,0)

m3 = (0,0,0,1,1,1,0)

Binary Search Over Hamming Distance

Oracle





ϕ_{query} :
Some property of the output
Ex: Some cells not selected

m1: True, m2: False

m3: True

Actively Learning Relevant Variables

$\boxed{Find\ U\ such\ that\ \phi_{explain}(V) \equiv \phi_{explain}(U)\ where\ |U| \ll |V|}$

Hamming Distance = 4

Assignments to V

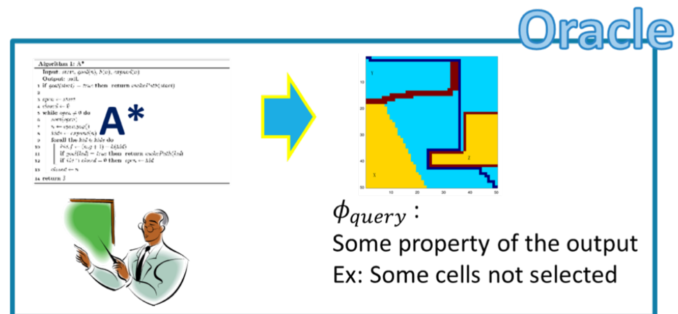
~~m1 = (0,0,0,1,1,0,1)~~

m2 = (0,0,1,1,0,1,0)

m3 = (0,0,0,1,1,1,0)

Hamming Distance = 2

Binary Search Over Hamming Distance



~~m1: True~~, m2: False

m3: True

Actively Learning Relevant Variables

$$\boxed{Find\ U\ such\ that\ \phi_{explain}(\mathrm{V}) \equiv \phi_{explain}(U)\ where\ |U| \ll |V|}$$

Hamming Distance = 2

Assignments to V

m2 = (0,0,1,1,0,1,0)

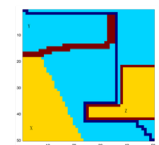
m3 = (0,0,0,1,1,1,0)

m4 = (0,0,1,1,1,1,0)

Hamming Distance = 1

Binary Search Over Hamming Distance

Oracle



ϕ_{query} :
Some property of the output
Ex: Some cells not selected

m2: False, m3: True

m4: True

Actively Learning Relevant Variables

$\boxed{Find\ U\ such\ that\ \phi_{explain}(V) \equiv \phi_{explain}(U)\ where\ |U| \ll |V|}$

Hamming Distance = 2

Assignments to V

m2 = (0,0,1,1,0,1,0)

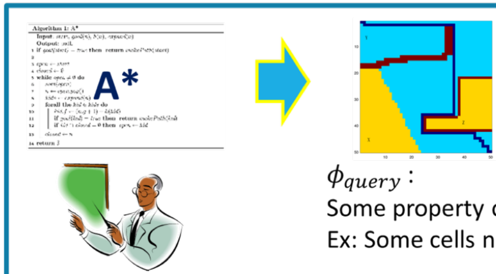
~~m3 = (0,0,0,1,1,1,0)~~

m4 = (0,0,1,1,1,1,0)

Hamming Distance = 1

Binary Search Over Hamming Distance

Oracle



ϕ_{query} :
Some property of the output
Ex: Some cells not selected

m2: False, ~~m3: True~~

m4: True

Actively Learning Relevant Variables

$$\boxed{Find\ U\ such\ that\ \phi_{explain}(V) \equiv \phi_{explain}(U)\ where\ |U| \ll |V|}$$

Hamming Distance = 1

Assignments to V

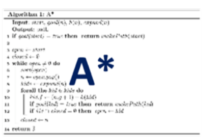
m2 = (0,0,1,1,$\boxed{0}$,1,0)

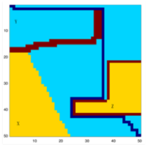
m4 = (0,0,1,1,$\boxed{1}$,1,0)

Fifth variable v_5 is relevant !!

Binary Search Over Hamming Distance

Oracle





ϕ_{query} :

Some property of the output

Ex: Some cells not selected

m2: False, m4: True

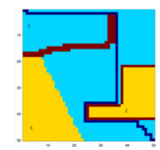
Actively Learning Relevant Variables

$$\boxed{Find\ U\ such\ that\ \phi_{explain}(\mathrm{V}) \equiv \phi_{explain}(U)\ where\ |U| \ll |V|}$$

Repeat to find all relevant variables

Binary Search Over Hamming Distance

Oracle



ϕ_{query} :
Some property of the output
Ex: Some cells not selected

m2: False, m4: True

Actively Learning Relevant Variables

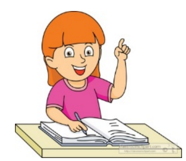
$\boxed{Find\ U\ such\ that\ \phi_{explain}(\mathrm{V}) \equiv \phi_{explain}(U)\ where\ |U| \ll |V|}$

For each assignment to relevant variables

$2^{|U|}$

Random Sample Till Oracle differs

$ln(1/(1-\kappa))$



Binary Search Over Hamming Distance

$ln(|V|)$

Relevant variables of $\boldsymbol{\phi_{explain}}$ found with confidence $\boldsymbol{\kappa}$ in

$$\mathbf{2^{|U|}}\boldsymbol{ln}(|\boldsymbol{V}|/(\mathbf{1}-\boldsymbol{\kappa}))$$

Actively Learning Boolean Formula

$\boxed{Find\ U\ such\ that\ \phi_{explain}(\mathrm{V}) \equiv \phi_{explain}(U)\ where\ |U| \ll |V|}$

Used distinguishing example based approach from ICSE’10

Susmit Jha, Sumit Gulwani, Sanjit A Seshia, and Ashish Tiwari. Oracle-guided component-based program synthesis. In *2010 ACM/IEEE 32nd International Conference on Software Engineering*, volume 1, pages 215–224. IEEE, 2010.

Build Truth Table for the relevant variables U

Scales to ~200 variables

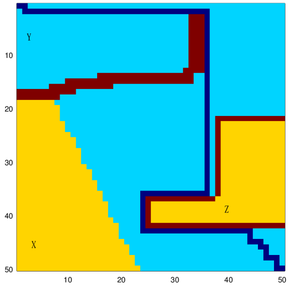
$Worst\ Case$: $2^{|U|}$

$\boldsymbol{\phi_{explain}}$ **found with confidence $\boldsymbol{\kappa}$ in**

$$\mathbf{O}(\mathbf{2}^{|\mathrm{U}|}\boldsymbol{ln}(|\boldsymbol{V}|/(\mathbf{1}-\boldsymbol{\kappa})))$$

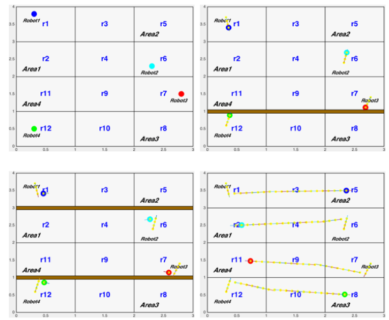
A PAC Learning Framework

Experiments



A* Planning
|V| = 2500
|U| <= 4
Runtime < 3 minutes

10^153



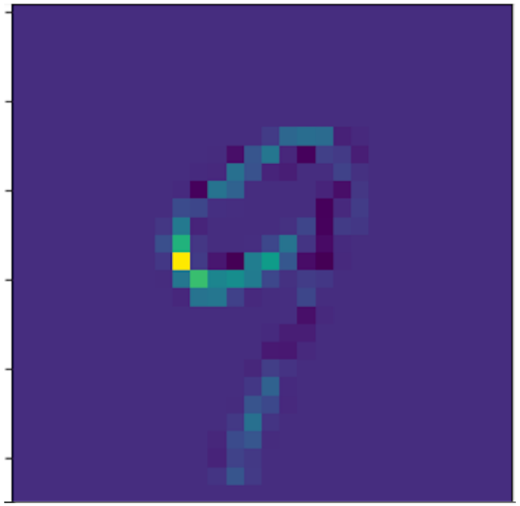
Reactive Exploration Strategy
|V| = 96
|U| <= 2
Runtime < 5 seconds

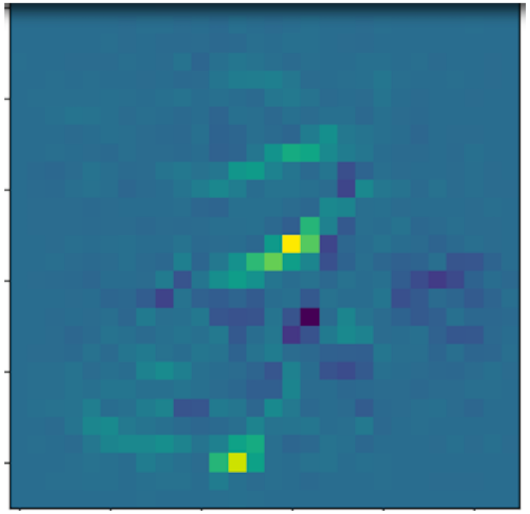
10^28

Image Classification: MNIST

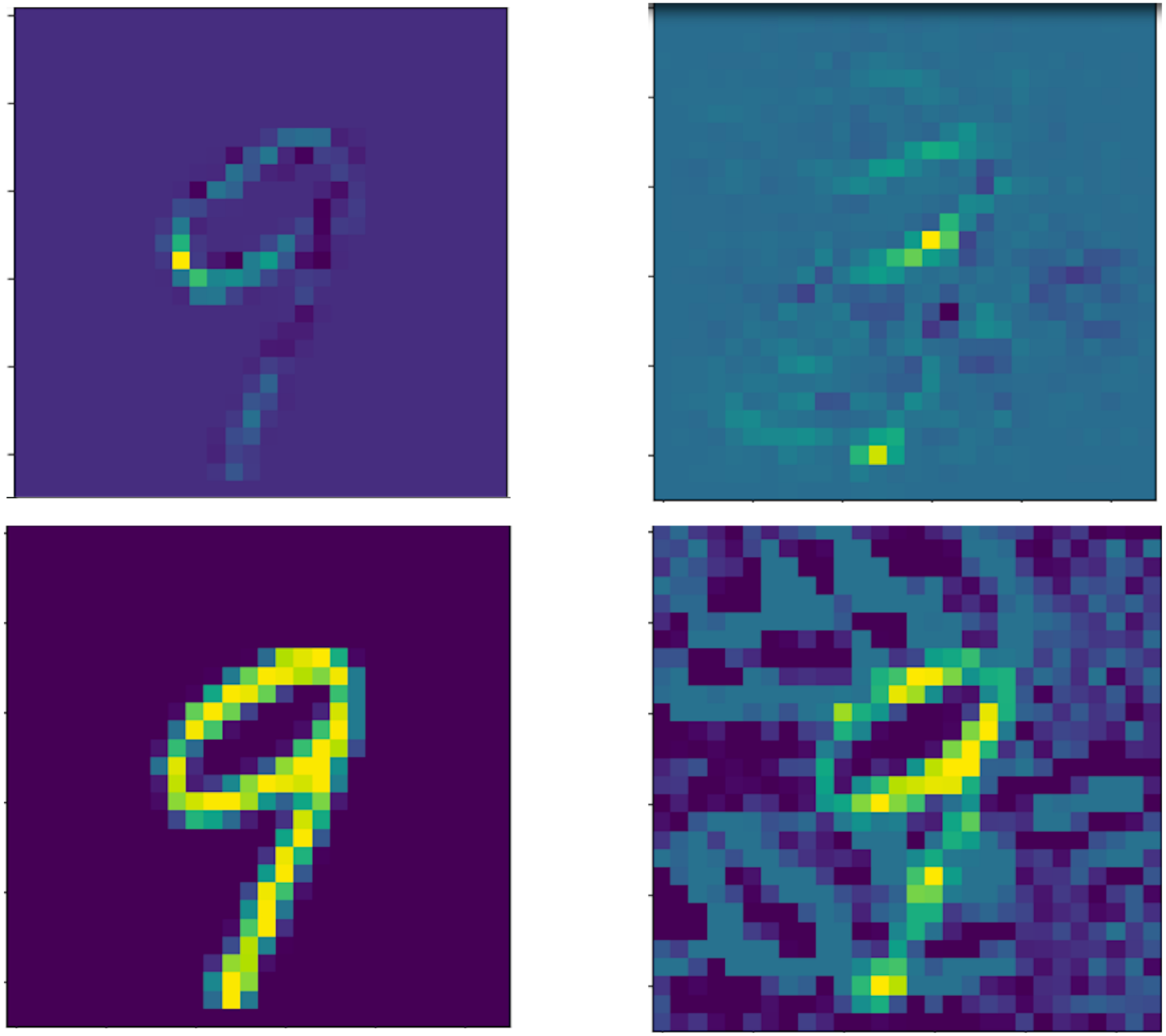
Image Classification: ImageNet with Carlini-Wagner Adversarial Attacks

Experiments

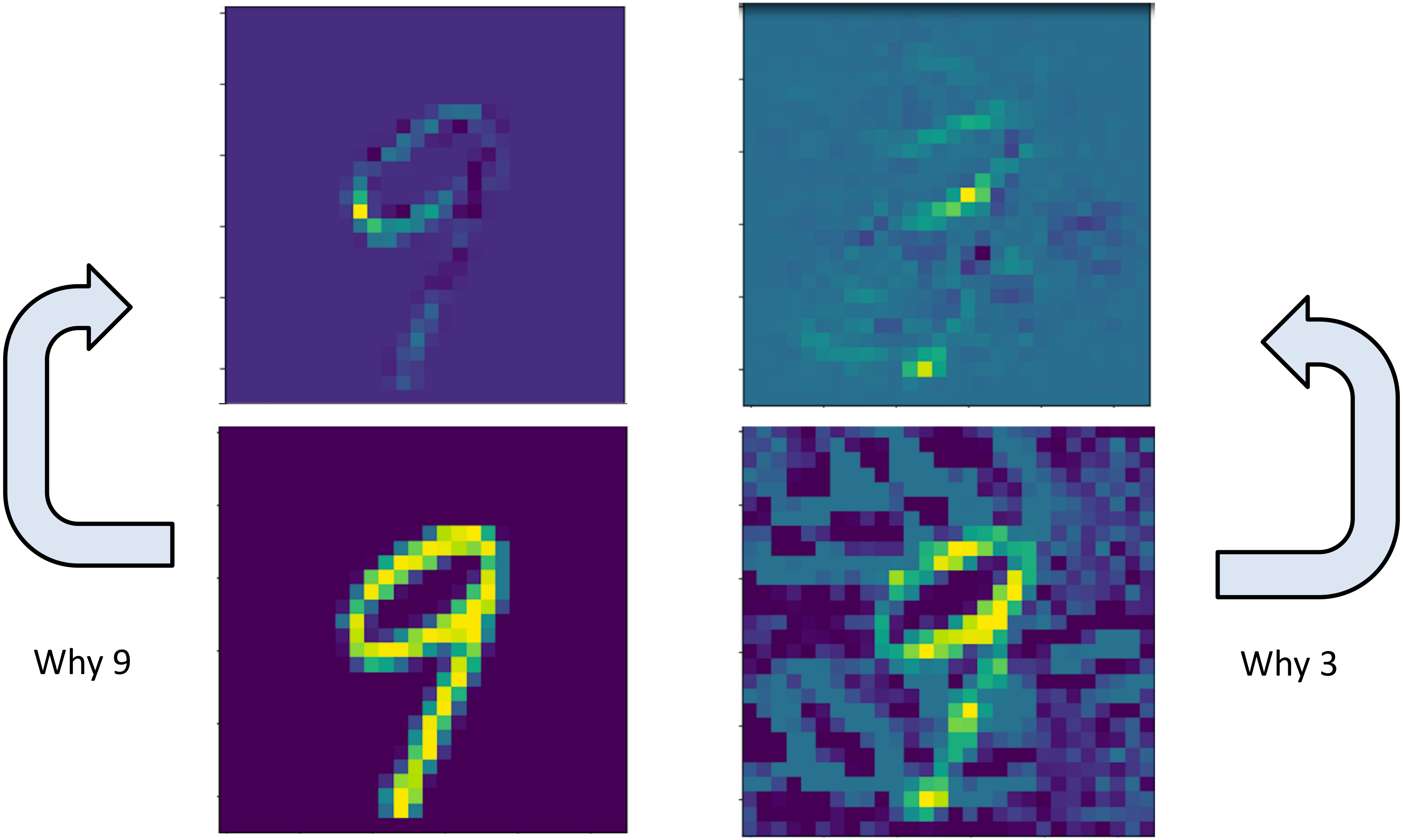




Experiments



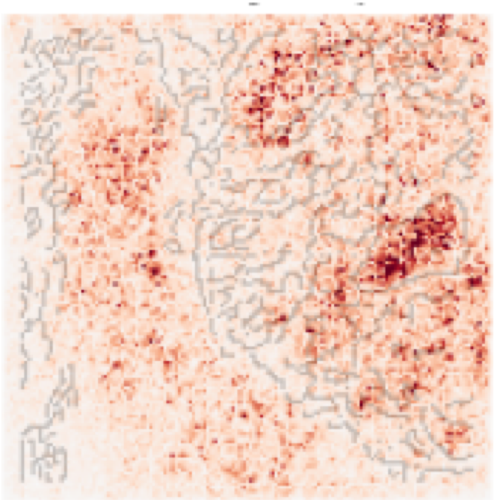
Experiments



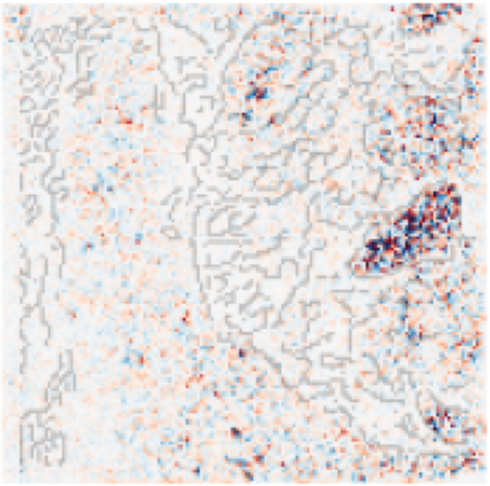
Why not just do sensitivity analysis?

Why not just do sensitivity analysis?

Why not just do sensitivity analysis?



Sensitivity (IG)

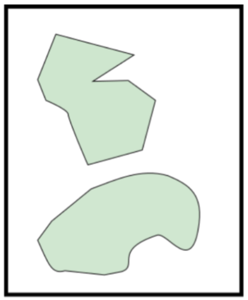


Sparse Boolean
Formula Learning

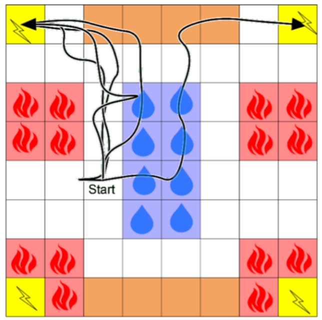
Learning Temporal Logic Properties from Noisy Time Traces

Bernoulli Distribution

$$\Pr\Big(\text{Specification} \;\Big|\; \text{Demonstrations} \Big) \propto e^{D_{KL}\left(\mathcal{B}(\overline{\varphi}) \,\|\, \mathcal{B}(\widehat{\varphi})\right)}$$



Specification



Demonstrations

Satisfaction probability for Alice given dynamics

Satisfaction probability given uniformly random actions

- Composable
- Resilient to changes in task context
- Interpretable
- Can leverage formal methods tools

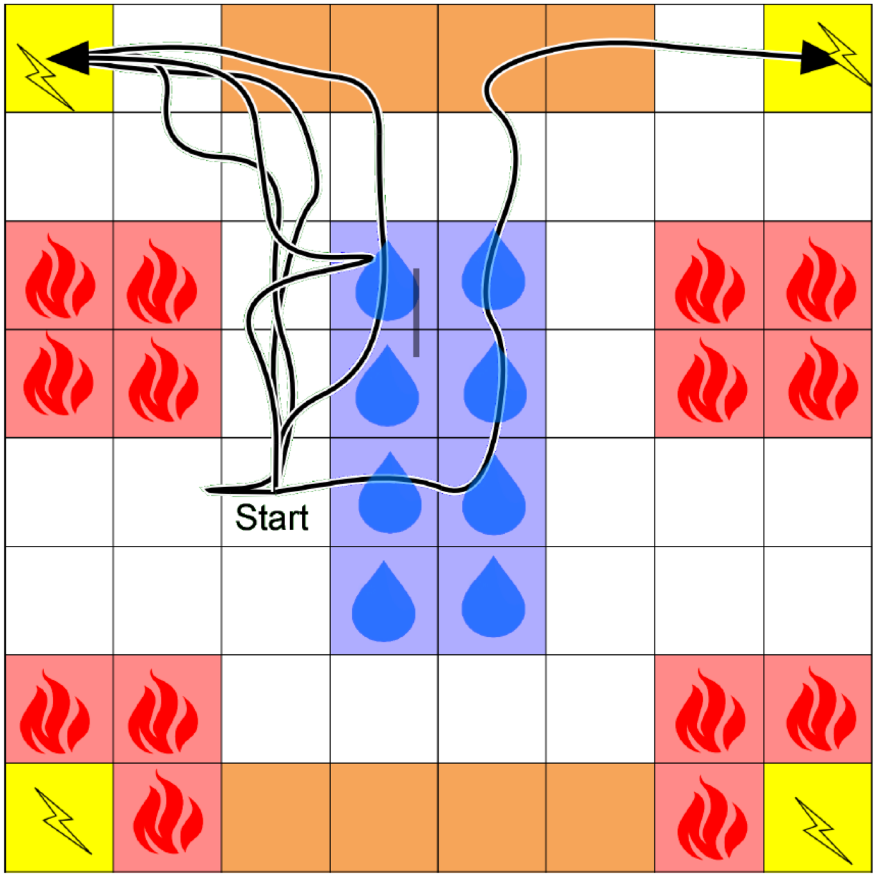
Marcell Vazquez-Chanlatte, Susmit Jha , Ashish Tiwari, Mark K. Ho and Sanjit A. Seshia. Learning Task Specifications from Demonstrations. NeurIPS, 2018

Communicating Using Demonstrations: More involved example

1. Avoid fire (red).

2. Eventually Recharge (yellow).

3. If you touch the water (blue) then dry off (brown) before recharging (yellow).



Temporal Logic Specification

H: Historically
O: Once
S: Since

$$\left(H\neg red \wedge O\ yellow\right) \wedge H\left(\left(yellow \wedge O\ blue\right) \Rightarrow \left(\neg blue\ S\ brown\right)\right)$$

Interpretability / Explanation Generation in TRINITY

- **Inferring and Conveying Intentionality: Beyond Numerical Rewards to Logical Intentions.** Susmit Jha and John Rushby.
  AAAI Spring Symposium, Towards Conscious AI Systems, 2019

- **Learning Task Specifications from Demonstrations.** Marcell Vazquez-Chanlatte, Susmit Jha, Ashish Tiwari, Mark K. Ho and Sanjit A. Seshia.
  Neural Information Processing Systems (NeurIPS), 2018

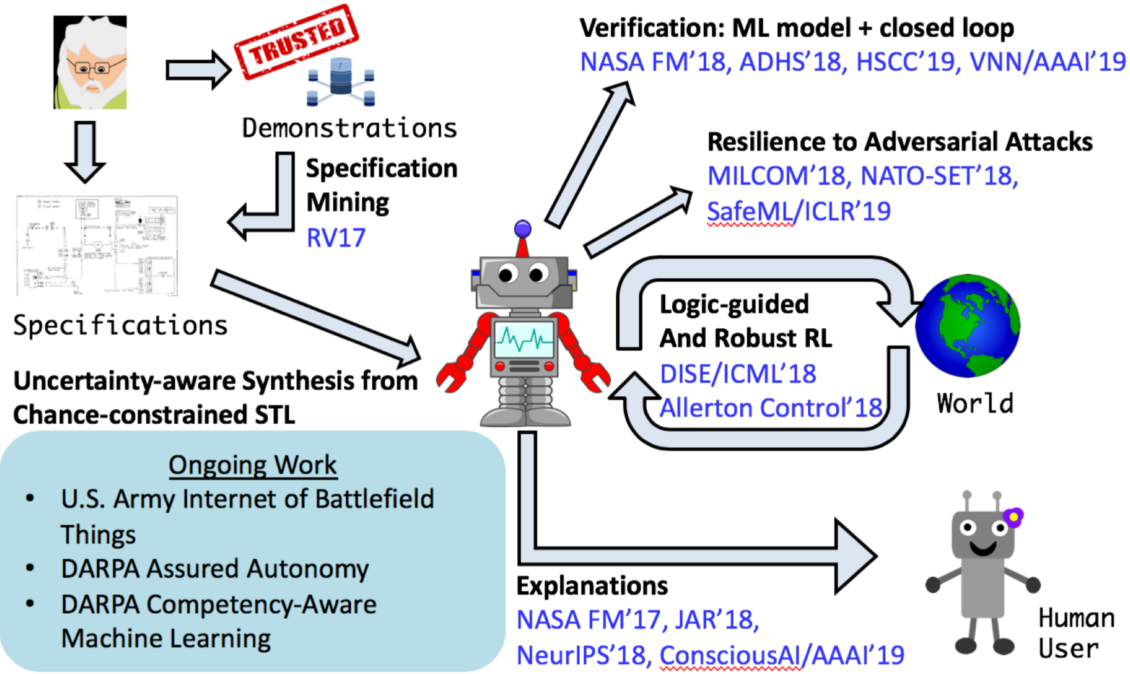
- **Explaining AI Decisions Using Efficient Methods for Learning Sparse Boolean Formulae.** Susmit Jha, Tuhin Sahai, Vasumathi Raman, Alessandro Pinto and Michael Francis.
  Journal of Automated Reasoning, 2018

- **On Learning Sparse Boolean Formulae For Explaining AI Decisions.** Susmit Jha, Vasumathi Raman, Alessandro Pinto, Tuhin Sahai, and Michael Francis.
  NASA Formal Methods (NFM), 2017

Thanks!

If you are interested in building *trusted, resilient and interpretable* AI, please contact me with your CV if you are interested.

TRINITY @ SRI



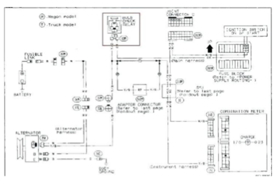
Co-travelers (Present and Past):
Brian Burns, Margaret Chapman, Ajay Divakaran, Sauradeep Dutta, Michael Francis, Mark K. Ho, Uyeong Jang, Brian Jalaian, Somesh Jha, Patrick Lincoln, Alessandro Pinto, Vasu Raman, John Rushby, Dorsa Sadigh, Sriram Sankaranarayanan, Sanjit A. Seshia, Natarajan Shankar, Ashish Tiwari, Claire Tomlin, Marcell Vazquez-Chanlatte, Gunjan Verma

Funding sources (Present and Past):
DARPA, US Army Research Laboratory, National Science Foundation

TRINITY @ SRI

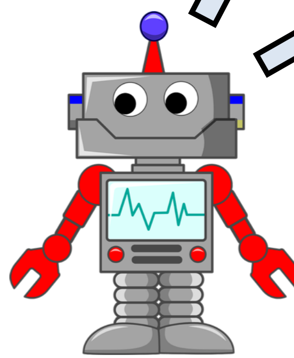
Demonstrations

Specification Mining
RV17



Specifications

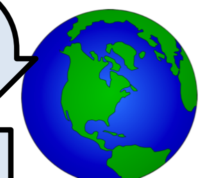
Uncertainty-aware Synthesis from Chance-constrained STL



Verification: ML model + closed loop
NASA FM'18, ADHS'18, HSCC'19, VNN/AAAI'19

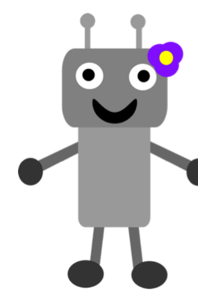
Resilience to Adversarial Attacks
MILCOM'18, NATO-SET'18, SafeML/ICLR'19

Logic-guided And Robust RL
DISE/ICML'18
Allerton Control'18



World

Explanations
NASA FM'17, JAR'18, NeurIPS'18, ConsciousAI/AAAI'19



Human User

Ongoing Work

- U.S. Army Internet of Battlefield Things
- DARPA Assured Autonomy
- DARPA Competency-Aware Machine Learning